February 2020

OBJECTIVE VOICE QUALITY PREDICTION FOR IP NETWORKS IN REAL TIME VOICE OVER INTERNET PROTOCOL APPLICATIONS

Hua Gao

Xinmin Yan

OBJECTIVE VOICE QUALITY PREDICTION FOR IP NETWORKS IN REAL TIME VOICE OVER INTERNET PROTOCOL APPLICATIONS

AUTHORS:

Hua Gao

Xinmin Yan

ABSTRACT

Presented herein is a system for monitoring/predicting objective voice quality in Voice Over Internet Protocol (VoIP) applications. The system is configured to , collect training datasets and generate online-training prediction models of voice quality.

DETAILED DESCRIPTION

Quality-of-Service (QoS) diagnostics is very important in current online products. Monitoring of voice quality, in particular, is challenging due to network variability characteristics and speech complexity. As such, proposed herein is a system that is configured to monitor and predict objective voice quality of VoIP, while collecting training datasets and online-training prediction models of voice quality. Figure 1, is a schematic block diagram of an example system, in accordance with the techniques presented herein.

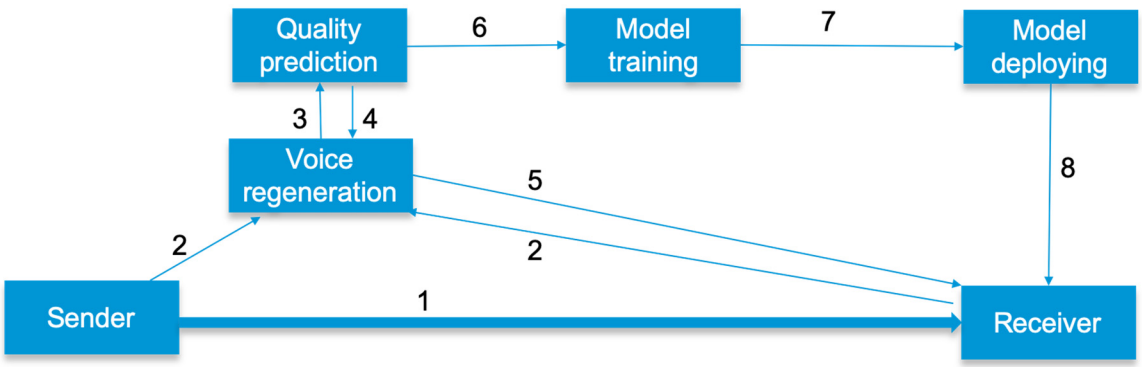

Figure 1

In the example of Figure 1, the sender is a VoIP transmitter (VoIP TX) and the receiver is a VoIP receiver (VoIP RX). The voice regeneration service is configured to regenerate sent/received voice data based on information from the sender/receiver. The quality prediction service is configured to calculate a MOS score with POLQA / PESQ tools. The model training service is configured to collective client encoder/decoder information and MOS score as training sets and generate online training models. The model deploying service is configured to deploy a latest prediction model to client.

In on example, a VoIP stream is sent from a sender to a receiver. The sender will select one part of stream as an evaluation sequence, will inform the receiver of the beginning packet and ending packet in the sequence, and record regeneration information, which can be uploaded to the voice regeneration service. The receiver will record regeneration information and upload to the voice regeneration service.

Information for regenerating voice or representations of VoIP service can include raw data of voice, network conditions (e.g., packet loss, delay and jitter, etc,), codecs, transmission information (e.g., retransmission, RTP fec, etc.), packets sequences and power, operation sequence of decoding (e.g., decoding packet time, PLC, in-band fec, decoder drop packet), etc.

The voice regeneration service will regenerate voice according to the received regeneration information. The voice regeneration service could, for example, regenerate the original voice in VoIP or simulation voice with a same behavior for security policies. If the voice regeneration and quality prediction services are located in a sender client, then original voice is preferred. The regenerated sender voice, as reference, and the receiver voice, regenerated as distortion speech, are uploaded to the quality prediction service.

Thereafter, a MOS score will be returned to the voice regeneration service and then the MOS score will be returned to the receiver client (correction to the local prediction of quality). All information and the MOS score will be uploaded to the model training service as a training set. A model can then be trained and/or updated for deploying services. The services can then be deployed/redeployed with new model coefficients.

The sender will select the voice period, which will be used to calculate MOS, and collect the information associated therewith. The PCM of the voice is the simplest information, but it may causes security/privacy problems. As such, the preferred

information is information which can represent the significant voice features, while preserving the users' privacy. For example, the energy envelop of the voice may be used instead of the PCM. Figure 2, below, illustrates sender information structure, where P(1) is the information of the first package of the voice that is selected. P(1) may include sequence number, timestamp, the energy of package, and VAD value.

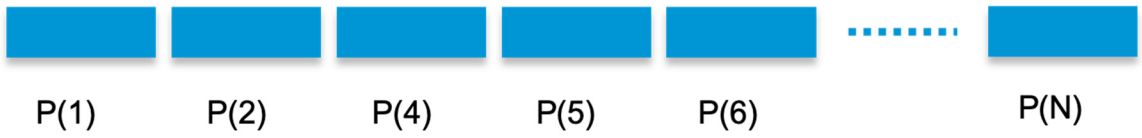

Figure 2

As noted, the receiver will record the regeneration information. The regeneration information is the behavior of the decoder processing the packets. For example , P(1) and P(2) may be received and decoded normally, while P(3) is delayed. Therefore, PLC or CNG packets were inserted into the decoded speech. P(4) was delayed too long, but Rs-FEC recovered P(4). However, P(4) was dropped. P(5) was lost by the network, then recovered by P(6), and so on. Figure 3, below, illustrates that the regeneration information structure consists of decoding time, packet length, timestamps related to P(1), packet type (normal, PLC, CNG, RS-FEC, Inband-FEC, etc.), packet number (normal packets, FEC packets), playback speed.

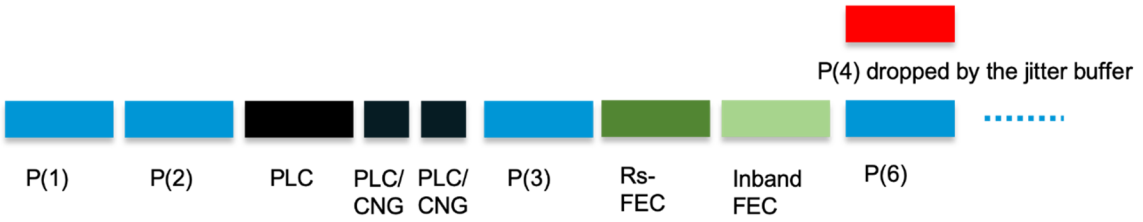

Figure 3

Figure 4, below, illustrates aspects of the voice regeneration and quality prediction. In particular, the voice sample dataset contains many voice pieces and the voice generator

will compose the voice pieces from the dataset. The composited voice will have a similar energy envelop as recorded by the sender information and is the reference speech of POLQA scored calculation. The composited voice will be encoded and decoded following the receiver behavior, where the decoded voice is the distorted speech.

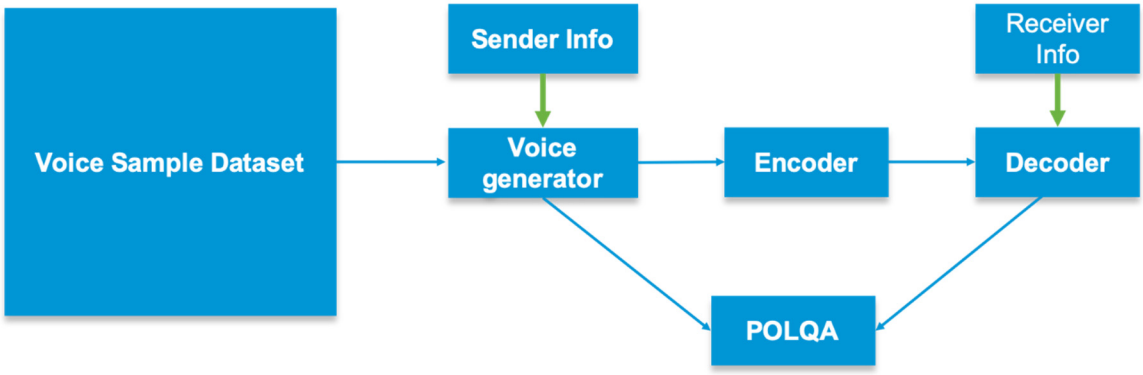

Figure 4

The purpose of training an artificial intelligent voice quality prediction model is to replace E-Model (ITU-T G.107). The model can calculate the estimation of received voice quality, which can be used to notify users of the audio service quality in real-time. In on example, the model uses the pare of <Receiver Info, POLQA score> as the training set. A linear regression model or support vector regression model could also be used as the training model.